Regularized LTI System Identification with Multiple Regularization Matrix

Chen, Tianshi; Andersen, Martin S.; Mu, Biqiang; Yin, Feng; Ljung, Lennart; Qin, S. Joe

Published in:
I F A C Workshop Series

Link to article, DOI:
10.1016/j.ifacol.2018.09.121

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 51-15 (2018) 180–185

Regularized LTI System Identification with Multiple Regularization Matrix ⋆

Tianshi Chen *, **Martin S. Andersen** †, **Biqiang Mu** ‡, **Feng Yin** *, **Lennart Ljung** ‡, **S. Joe Qin** §

* *School of Science and Engineering and Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China*
† *Department of Applied Mathematics and Computer Science, Technical University of Denmark, Copenhagen, Denmark*
‡ *Department of Electrical Engineering, Linköping University, Linköping, Sweden*
§ *Department of Electrical Engineering and Department of Chemical Engineering and Materials Science, University of Southern California*

Abstract: Regularization methods with regularization matrix in quadratic form have received increasing attention. For those methods, the design and tuning of the regularization matrix are two key issues that are closely related. For systems with complicated dynamics, it would be preferable that the designed regularization matrix can bring the hyper-parameter estimation problem certain structure such that a locally optimal solution can be found efficiently. An example of this idea is to use the so-called multiple kernel Chen et al. (2014) for kernel-based regularization methods. In this paper, we propose to use the multiple regularization matrix for the filter-based regularization. Interestingly, the marginal likelihood maximization with the multiple regularization matrix is also a difference of convex programming problem, and a locally optimal solution could be found with sequential convex optimization techniques.

Keywords: System identification, regularization methods, sequential convex optimization.

1. INTRODUCTION

The traditional system identification is to construct mathematical models based on the measured input-output data, see e.g., Ljung (1999). Recently, there have been increasing research interests in the system identification community to further integrate the prior knowledge in the construction of the mathematical models. To differentiate them from the traditional system identification, they are referred to as the regularized system identification in this paper. The idea to utilize the prior knowledge together with the input-output data is by no means new, see e.g., (Ljung, 1999, p. 15). There are different ways to integrate the prior knowledge of the system to be identified into the model estimation, e.g., by using heuristics, Bayesian methods or regularization methods. However, no clear trend was formed in the system identification community until recently, see e.g., Pillonetto and Nicolao (2010); Latarie and Chen (2016); Pillonetto and Chiuso (2015); Chen et al. (2012, 2014); Marconato et al. (2016); Chen (2018); Mu et al. (2018b); Mu and Chen (2018); Chen (2019); Chen and Pillonetto (2018); Mu et al. (2017, 2018a); Hong et al. (2018); Prando et al. (2017); Zorzi and Chiuso (2017); see Pillonetto et al. (2014) for a survey and Chiuso (2016) for a review of regularization methods. The major obstacle is that it was unclear how to embed the prior knowledge of a system to be identified into the regularization. The intriguing finding disclosed by Pillonetto and Nicolao (2010); Chen et al. (2012, 2014); Chen (2018); Marconato et al. (2016); Zorzi and Chiuso (2017) is that when considering impulse response estimation problem of linear time invariant (LTI) systems, it is possible to design a regularization in quadratic form to embed the prior knowledge of the impulse response to be identified.

For the regularization in quadratic form, there exist different ways to design the regularization matrix. One way is to design the regularization matrix through a positive semidefinite kernel Chen et al. (2012, 2014); Chen (2018). Accordingly, this kind of regularization is referred to as the kernel-based regularization. In particular, two systematic ways to design the kernels are proposed in Chen (2018): one way is from a machine learning perspective and another way is from a system theory perspective. The machine learning perspective treats the impulse response as a function and the prior knowledge could be about the decay and varying rate of the impulse response. Then we can design the so-called amplitude modulated locally stationary (AMLS) kernel, which is a multiplication of a rank-

⋆ This work was supported by the Thousand Youth Talents Plan of China, the general projects funded by NSFC under contract No. 61773329 and 61603379, the Shenzhen research projects funded by the Shenzhen Science and Technology Innovation Council under contract No. Ji-20170189 and Ji-20160207, the President's grant under contract No. PF. 01.000249 and the Start-up grant under contract No. 2014.0003.23 funded by the Chinese University of Hong Kong, Shenzhen, as well as by a research grant for junior researchers funded by Swedish Research Council under contract No. 2014-5894, the National Key Basic Research Program of China (973 Program) under contract No. 2014CB845301, the President Fund of AMSS, CAS under contract No. 2015-hwyxqnrc-mbq.

2405-8963 © 2018, IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2018.09.121

1 kernel and a stationary kernel, parameterized to account for the decay and varying rate of the impulse response, respectively. The system theory perspective associates the impulse with an LTI system and the prior knowledge could be that the system is stable and may be overdamped, underdamped, have multiple distinct time constants and etc. Then we can design the so-called simulation induced (SI) kernel using the multiplicative uncertainty configuration from robust control theory. In particular, the nominal model is used to embed the prior knowledge, the uncertainty is assumed to be stable and finally the system is simulated with an impulsive input to get the SI kernel. A recent contribution along this way is Zorzi and Chiuso (2017), where the harmonic analysis of AMLS kernels is provided and more general kernels are designed.

Another way to design the regularization matrix is through a filter matrix Marconato et al. (2016), which is motivated by the special structure of the diagonal correlated (DC) kernel Chen et al. (2016); Marconato et al. (2016); Carli et al. (2017). The filter matrix is an upper triangular matrix whose rows are coefficients of filters designed based on the prior knowledge, which could be that the underlying system is low-pass, high-pass, band-pass, and etc. Moreover, to guarantee stability, those filters differ in their gains: filters associated with rows with high number will have higher gains. The filters are parameterized by the hyperparameter, which could be the order of the filter, the cut-off frequencies, the decay rate and the scaling factor.

It is worth to note that the issue of the regularization matrix design is closely related with the issue of hyperparameter estimation Chen et al. (2014). If the prior knowledge is that the system to be identified has complicated dynamics, e.g., the system has multiple time constants, multiple cut-off frequencies, it will make sense to design the regularization matrix with complicated structure. However, a regularization matrix with complicated structure would make the nonconvex hyperparameter estimation problem hard to solve. So when designing the regularization matrix, we should also consider whether the designed regularization matrix can bring the hyperparameter estimation problem certain structure such that a locally optimal solution could be found efficiently. An example of this idea is to use the so-called multiple kernel Chen et al. (2014) for kernel-based regularization methods. In this paper, we extend the idea of Chen et al. (2014) and propose to use the multiple regularization matrix for the filter-based regularization method. Interestingly, the marginal likelihood maximization with the multiple regularization matrix can also be expressed as a difference of convex programming (DCP) problem, and a locally optimal solution could thus be found efficiently with sequential convex optimization techniques.

2. SYSTEM IDENTIFICATION

Consider a discrete-time stable and causal LTI system

$$y(t) = G(q)u(t) + v(t), \quad t = 1, 2, \cdots, N, \tag{1}$$

where t is the time index, N is the number of observations, $G(q)$ is the transfer function of the LTI system with q being the forward shift operator, and $y(t)$, $u(t)$, $v(t) \in \mathbb{R}$ are the measured output, input, disturbance at time instant t, respectively. Here it is assumed that the disturbance $v(t)$ is a zero mean white noise with variance $\sigma^2 > 0$ and moreover, independent of the input $u(t)$ for $t \in \mathbb{N}$. *The traditional system identification problem is to estimate* $G(q)$ *as well as possible based on the data* $\{y(t), u(t)\}_{t=1}^N$.

The prediction error/maximum likelihood method (M-L/PEM) is the traditional method for LTI system identification. It first proposes a parametric model structure, e.g., a parameterization of $G(q)$ and then derives the model estimate by minimizing the prediction error criterion.

The simplest model structure is perhaps the finite impulse response (FIR) model which takes the following form:

$$G(q) = \sum_{k=1}^{n} g(k)q^{-1}. \tag{2}$$

In this way, the estimation of $G(q)$ becomes to the estimation of an FIR model of system (1) based on the data $\{y(t), u(t)\}_{t=1}^N$:

$$y(t) = \sum_{k=1}^{n} g(k)u(t-k) + v(t), \quad t = 1, 2, \cdots, N, \tag{3}$$

where it is common to assume that $N \gg n$.

Clearly, $y(t)$ with $t \in \mathbb{N}$ depends on $u(t-1), \ldots, u(t-n)$, which may not be known. There are different ways to deal with the unknown input $u(t)$ with $t = 0, -1, \ldots, 1-n$. Accordingly, there are different ways to rewrite the FIR model (3) in a matrix form:

$$Y_M = \Phi_M \theta + V_M, \tag{4}$$

where $Y_M \in \mathbb{R}^M, \Phi_M \in \mathbb{R}^{M \times n}, V_M \in \mathbb{R}^M$ and

$$\theta = [\, g(1)\ g(2)\ \cdots\ g(n)\,]^T, \tag{5}$$

which is the parameter of the FIR model (2) and simply called the impulse response in the sequel. For instance, see e.g., Chen et al. (2012), if we choose not to use the unknown input $u(t)$ with $t = 0, -1, \ldots, 1-n$ and assume that $N > n$, then $M = N - n$,

$$Y_M = Y_{N-n} = [\, y(n+1)\ y(n+2)\ \cdots\ y(N)\,]^T, \tag{6}$$

and the expressions of Φ_M and V_M can be figured out in a straightforward way.

For the case where the disturbance $v(t)$ in (1) is Gaussian and the FIR model (2) is used, the PEM/ML becomes the least squares method

$$\hat{\theta}^{LS} = \arg\min_{\theta} \|Y_M - \Phi_M \theta\|_2^2, \tag{7}$$

where $\|\cdot\|_2$ is the Euclidean norm. A disadvantage of $\hat{\theta}^{LS}$ is that it may be subject to high variance for large n. The major difficulty of the traditional PEM/ML is to choose a suitable model complexity, i.e., the model order, which is often done by applying cross validation or complexity criteria such as AIC, BIC, or model validation techniques, see e.g., Ljung (1999). However, these classical techniques are sometimes not as reliable as expected, see e.g., Chen et al. (2012); Pillonetto et al. (2014).

3. REGULARIZED SYSTEM IDENTIFICATION

In practice, there may exist prior knowledge of the system (1) besides the data $\{y(t), u(t)\}_{t=1}^N$. Interestingly, the question of how to make use of such prior knowledge for a better estimate of $G(q)$ was not systematically investigated

in the system identification community until very recently. To differentiate it from the traditional system identification, it is referred to as *regularized system identification* aiming to estimate $G(q)$ as well as possible based on both the data $\{y(t), u(t)\}_{t=1}^N$ and the prior knowledge of $G(q)$.

There are different ways to integrate the prior knowledge of $G(q)$ into the estimation of $G(q)$, e.g., by using heuristics, Bayesian methods and regularization methods. When the prior knowledge of $G(q)$ is about the impulse response θ, it is possible to embed the prior knowledge in a suitably defined regularization matrix Chen et al. (2012). To be specific, it is possible to design a suitable regularization matrix such that the impulse response θ is estimated by the so-called regularized least squares method:

$$\hat{\theta}^R = \arg\min_{\theta} \|Y_M - \Phi_M\theta\|_2^2 + \gamma\theta^T D\theta, \tag{8}$$

where $D \in \mathbb{R}^{n\times n}$ is a positive semidefinite matrix called the regularization matrix, and $\gamma > 0$ is the regularization parameter used to balance the trade-off between the adherence to the data and the penalty on the model complexity.

To find a suitable regularization matrix D consists of two steps. Based on the prior knowledge, the first step is to propose a parameterization of D with a parameter vector $\eta \in \Omega \in \mathbb{R}^p$, called the hyper-parameter, where Ω is a set in which we search for a suitable hyper-parameter η. For the parameterized regularization matrix $D(\eta)$, the second step is to estimate η based on the data $\{y(t), u(t)\}_{t=1}^N$.

3.1 Kernel-based and Filter-based Regularization

There are different ways to design the regularization matrix D. One way is to design D through a positive semidefinite kernel function designed based on the prior knowledge, see e.g., Chen (2018). Recall that a function $k : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ is called a positive semidefinite kernel (simply called a kernel hereafter), if it is symmetric and satisfies $\sum_{i,j=1}^m a_i a_j k(x_i, x_j) \geq 0$ for any $m \in \mathbb{N}$, $\{x_1, \cdots, x_m\} \subset \mathbb{N}$ and $\{a_1, ..., a_m\} \subset \mathbb{R}$. Now let $k(t, s; \eta)$ be a kernel parameterized by a hyper-parameter $\eta \in \Omega \in \mathbb{R}^p$. Then one can choose $D(\eta) = (K(\eta))^{-1}$ with $K \in \mathbb{R}^{n\times n}$ and $K_{i,j} = k(i, j; \eta)$, where K is called the kernel matrix. Since this kind of regularization matrices relies on a kernel, it is referred to as the kernel-based regularization method in Chen et al. (2012, 2014); Chen (2018). Accordingly, the regularized least squares criterion (8) becomes

$$\hat{\theta}^R = \arg\min_{\theta} \|Y_M - \Phi_M\theta\|_2^2 + \gamma\theta^T (K(\eta))^{-1}\theta. \tag{9}$$

As is well-known, if we let $\gamma = \sigma^2$ and assume that $\theta \sim \mathcal{N}(0, K(\eta))$ and $v(t)$ is Gaussian distributed, then $\hat{\theta}^R$ is the same as the maximum a posteriori estimate

$$\hat{\theta}^{MAP} = \arg\max_{\theta} p(\theta|Y_M). \tag{10}$$

In this case, the kernel matrix K is interpreted as the prior covariance matrix of θ. So far many kernels have been introduced, e.g., the stable spline (SS) kernel Pillonetto and Nicolao (2010) and the diagonal correlated (DC) kernel Chen et al. (2012), and the multiple kernel Chen et al. (2014); see Chen (2018) for two systematic ways to design more general kernels.

Motivated by the special structure of the DC kernel Chen et al. (2016); Marconato et al. (2016); Carli et al. (2017), another way to design D through a filter matrix was introduced recently in Marconato et al. (2016). Recall that a filter matrix F is an upper triangular matrix whose rows are coefficients of filters designed based on the prior knowledge and moreover, those filters only differ in their gains: filters associated with rows with high number will have higher gains. The filters are parameterized by the hyperparameter $\eta \in \Omega \in \mathbb{R}^p$, which could be the order of the filter, the cut-off frequencies, the decay rate and the scaling factor. Then one can choose $D(\eta) = (F(\eta))^T F(\eta)$ and as a result, $F(\eta)\theta$ can be interpreted as the response of the filter with the impulse response θ as the input. Accordingly, the regularized least squares (8) becomes

$$\hat{\theta}^R = \arg\min_{\theta} \|Y_M - \Phi_M\theta\|_2^2 + \gamma\|F(\eta)\theta\|_2^2. \tag{11}$$

It is clear to see that only the frequency components of the impulse response θ which pass the filters will be penalized in the cost function. Moreover, if high-pass filters are chosen, then the high frequency components of the impulse response will be penalized, leading to smooth impulse responses; if the filters associate with rows with high number have higher gains, the impulse response coefficients in the end are penalized more than the ones in the beginning, leading to decaying impulse responses.

3.2 Hyper-parameter Estimation

There are different ways to deal with the hyper-parameter estimation problem. So far the most widely used method is the empirical Bayes method, which embeds the regularization term in a Bayesian framework and then maximizes the marginal likelihood to get an estimate of η. More specifically, we let $\gamma = \sigma^2$ and assume that $D(\eta)$ is nonsingular, $\theta \sim \mathcal{N}(0, (D(\eta))^{-1})$, and $v(t)$ is Gaussian distributed. Then

$$\begin{aligned}\hat{\eta} &= \arg\max_{\eta\in\Omega} p(Y_M|\eta) \\ &= \arg\min_{\eta\in\Omega} Y_M^T\Sigma(\eta)^{-1}Y_M + \log\det\Sigma(\eta) \\ \Sigma(\eta) &= \Phi_M(D(\eta))^{-1}\Phi_M^T + \sigma^2 I_M\end{aligned} \tag{12}$$

Remark 3.1. When D is singular, the Moore-Penrose pseudo inverse D^+ of D should be used instead in (12). For convenience, we only consider the case where D is nonsingular in this paper.

4. MULTIPLE REGULARIZATION MATRIX BASED REGULARIZATION

Regardless of how the regularization is designed, there are some common difficulties for this regularization method. For instance, in order to model complicated systems, the regularization should have complicated structure. However, the complicated structure will make the nonconvex hyperparameter estimation problem difficult to solve. So we should consider whether the designed regularization can bring the hyper-parameter estimation problem certain structure such that a locally optimal solution could be found efficiently when we design the regularization.

An example of this idea is the use of multiple kernel Chen et al. (2014) for the kernel-based regularization method. In particular, the multiple kernel matrix is a conic combination of some fixed kernel matrices K_i:

$$K(\eta)=\sum_{i=1}^{p}\eta_i K_i,\ \eta_1,\ldots,\eta_p\geq 0,\ \eta=[\,\eta_1\ \eta_2\ \cdots\ \eta_p\,], \tag{13}$$

where K_i, $i=1,\ldots,p$, can be instances of DC kernel matrices with different decay rates and correlation coefficients. The multiple kernel equips the kernel-based regularization method with a couple of features. For example, multiple kernels can better capture complicated dynamics than a single kernel. Moreover, the hyperparameter estimation problem by maximizing the marginal likelihood can be expressed as a difference of convex programming (DCP) problem. Thus, it is possible to find a locally optimal solution efficiently by using sequential convex optimization techniques.

This idea could be extended and multiple regularization matrix could be used for the filter-based regularization method. In particular, the multiple regularization matrix $D(\eta)$ is positive definite and a nonzero conic combination of some fixed regularization matrices:

$$\begin{aligned}D(\eta)&=\sum_{i=1}^{p}\eta_i D_i=\sum_{i=1}^{p}\eta_i F_i^T F_i,\ \eta_1,\ldots,\eta_p\geq 0,\\ D_i&=F_i^T F_i,\eta=[\eta_1,\cdots,\eta_p]^T,\end{aligned} \tag{14}$$

where the given filter matrices $F_i\in\mathbb{R}^{n\times n}$, $i=1,\ldots,p$, are nonsingular and can be instances of the filter matrix proposed in Marconato et al. (2016) with different decay rates (and possibly with different orders and different cut-off frequencies). Clearly, the remaining problem is whether or not the multiple regularization matrix brings the hyperparameter estimation certain structure such that a locally optimal solution could be found efficiently. Fortunately, the answer to this problem is positive.

Remark 4.1. We have assumed for convenience that F_i, $i=1,\ldots,n$ are nonsingular and square in (14), but this assumption can be weakened.

Proposition 4.1. Consider the marginal likelihood maximization problem (12) with the multiple regularization matrix (14). Then the marginal likelihood maximization problem (12) can be expressed as a DCP problem for nonzero η.

Proof: By using the matrix inversion lemma, we have

$$\begin{aligned}Y_M^T\Sigma(\eta)^{-1}Y_M&=\|Y_M\|_2^2/\sigma^2\\ &\quad-Y_M^T\Phi_M(D+\Phi_M^T\Phi_M/\sigma^2)^{-1}\Phi_M^T Y_M/\sigma^4.\end{aligned} \tag{15}$$

Then by Sylvester's determinant identity, we have

$$\begin{aligned}\log\det\Sigma(\eta)&=M\log\sigma^2+\log\det(\Phi_M D^{-1}\Phi_M^T/\sigma^2+I_M)\\ &=M\log\sigma^2+\log\det(D^{-1}\Phi_M^T\Phi_M/\sigma^2+I_n)\\ &=M\log\sigma^2-\log\det(D)+\log\det(D+\Phi_M^T\Phi_M/\sigma^2).\end{aligned} \tag{16}$$

Note that $-\log\det(D)$ is convex for $D>0$, $-Y_M^T\Phi_M(D+\Phi_M^T\Phi_M/\sigma^2)^{-1}\Phi_M^T Y_M/\sigma^4$ and $\log\det(D+\Phi_M^T\Phi_M/\sigma^2)$ are concave for $D>0$. Moreover, since D is affine in η, $-\log\det(D)$ is convex for nonzero η, $-Y_M^T\Phi_M(D+\Phi_M^T\Phi_M/\sigma^2)^{-1}\Phi_M^T Y_M/\sigma^4$ and $\log\det(D+\Phi_M^T\Phi_M/\sigma^2)$ are concave for nonzero η. This completes the proof. □

From now on, we let

$$f(\eta)=g(\eta)-h(\eta),\ \eta\in\Omega=\{\eta_i\geq 0,i=1,\ldots,p\} \tag{17}$$

$$g(\eta)=-\log\det(D)=-\log\det(\sum_{i=1}^{p}\eta_i F_i F_i^T), \tag{18}$$

$$\begin{aligned}h(\eta)&=Y_M^T\Phi_M(D+\Phi_M^T\Phi_M/\sigma^2)^{-1}\Phi_M^T Y_M/\sigma^4\\ &\quad-\log\det(D+\Phi_M^T\Phi_M/\sigma^2)-\|Y_M\|_2^2/\sigma^2-M\log\sigma^2\\ &=Y_M^T\Phi_M(\sum_{i=1}^{p}\eta_i F_i F_i^T+\Phi_M^T\Phi_M/\sigma^2)^{-1}\Phi_M^T Y_M/\sigma^4\\ &\quad-\log\det(\sum_{i=1}^{p}\eta_i F_i F_i^T+\Phi_M^T\Phi_M/\sigma^2)\\ &\quad-\|Y_M\|_2^2/\sigma^2-M\log\sigma^2.\end{aligned} \tag{19}$$

Then the maximization problem (12) with the multiple regularization matrix (14) can be written as follows:

$$\hat{\eta}=\arg\min_{\eta\in\Omega}f(\eta)=\arg\min_{\eta\in\Omega}g(\eta)-h(\eta). \tag{20}$$

It follows from Proposition 4.1 that $g(\eta)$ and $h(\eta)$ are convex for $\eta\geq 0$ and $\eta\neq 0$, respectively.

As discussed in Chen et al. (2014), a locally optimal solution can be found efficiently by using sequential convex optimization techniques, e.g., the majorization minimization (MM) algorithm Yuille and Rangarajan (2002); Hunter and Lange (2004). The idea of MM is to derive an iterative optimization scheme for $\text{minimize}_{\eta\in\Omega}f(\eta)$. At each iteration, a so-called majorization function $\bar{f}(\eta,\eta^{(k)})$ of $f(\eta)$ at $\eta^{(k)}\in\Omega$ is minimized:

$$\eta^{(k+1)}=\arg\min_{\eta\in\Omega}\bar{f}(\eta,\eta^{(k)}), \tag{21}$$

where $\bar{f}:\Omega\times\Omega\to\mathbb{R}$ satisfies $\bar{f}(\eta,\eta)=f(\eta)$ for $\eta\in\Omega$ and $f(\eta)\leq\bar{f}(\eta,z)$ for $\eta,z\in\Omega$. Clearly, (21) yields an iterative descent algorithm with $k=1,2,\cdots$.

The construction of a suitable majorization function is a key step for MM algorithms. For the DCP problem (20), there are different ways to construct the majorization function Hunter and Lange (2004). Here we choose to use the so-called linear majorization, i.e.,

$$\bar{f}(\eta,\eta^{(k)})=g(\eta)-h(\eta^{(k)})-\nabla h(\eta^{(k)})^T(\eta-\eta^{(k)}), \tag{22}$$

where $\nabla h(\eta^{(k)})$ is the gradient of $h(\eta)$ evaluated at $\eta=\eta^{(k)}$. For this particular choice of majorization function, the MM algorithm (21) is also referred to as "sequential convex optimization" or "the convex concave procedure" (CCCP) Yuille and Rangarajan (2002).

Now we let

$$\bar{\Sigma}(\eta)=\sum_{i=1}^{p}\eta_i F_i F_i^T+\Phi_M^T\Phi_M/\sigma^2. \tag{23}$$

Then the MM algorithm for the problem (12) with the multiple regularization matrix (14) can be summarized as follows: choose $\eta^{(0)}\in\Omega$, set $k=0$ and then go to the following iterative steps:

Step 1: For $i=1,\cdots,p$, we calculate $\nabla h(\eta^{(k)})$:

$$\begin{aligned}\nabla_{\eta_i}h(x)=-\text{Tr}\,\big(&\bar{\Sigma}(\eta)^{-1}(\Phi_M^T Y_M Y_M^T\Phi_M/\sigma^4\\ &+\bar{\Sigma}(\eta))\bar{\Sigma}(\eta)^{-1}F_i^T F_i\big)\end{aligned} \tag{24}$$

and then solve the convex optimization problem (21) and (22) to obtain $\eta^{(k+1)}$.

Step 2: Check if the optimality condition is satisfied. If satisfied, stop. If otherwise, set $k = k + 1$ and go to step 1.

Remark 4.2. In Marconato et al. (2016), the hyper-parameter estimation problem is tackled by using the k-fold cross validation Hastie et al. (2001) and it was mentioned in (Marconato et al., 2016, p. 200) that the implementation of the hyper-parameter tuning could be improved. Here, we tried the empirical Bayes method and moreover, we have showed that for the multiple regularization matrix, the empirical Bayes method can be expressed as a DCP problem for which a locally optimal solution could be found efficiently by using sequential convex optimization techniques.

Remark 4.3. In the multiple regularization matrix (14), the reason why we keep $D_i = F_i^T F_i$ is two fold. First, for the filter-based regularization, the filter matrix F_i is designed directly but not the regularization matrix D_i. Second, in the implementation, we often need the factorization F_i of D_i for efficient and reliable implementation.

5. NUMERICAL SIMULATION

In this section, we aim to test, on the one hand, the idea to use multiple regularization matrix for regularization methods, and on the other hand, a locally optimal solutions of the marginal likelihood maximization (12) with the multiple regularization matrix (14) can be found with sequential convex optimization techniques.

5.1 Test Systems and Data-bank

The method in (Chen et al., 2012, Section 2) is used to generate 100 test systems, each of which is a 30th order discrete-time LTI system. Each test system is then simulated with a white Gaussian noise input $u(t)$ and the corresponding output is referred to as the noise-free output. The noise-free output is then perturbed by an additive white Gaussian noise and the perturbed output is collected as the output $y(t)$ for each test system. The signal-to-noise ratio, i.e., the ratio between the variance of the noise-free output and the noise, is chosen to be equal to 1. In this way, we collect $N = 375$ data points, i.e., $y(t), u(t)$ with $t = 1, 2, \cdots, 375$.

5.2 Simulation Setup

For each test system and data set, we construct two regularized FIR models with order $n = 125$ by using the following two methods.

The first method is to use the kernel-based regularization method (9) with the tune-correlated (TC) kernel introduced in Chen et al. (2012)

$$k^{\mathrm{TC}}(i, j; \eta) = c \min\{\lambda^i, \lambda^j\},\ i, j = 1, \cdots, n, \qquad (25)$$
$$\eta = [c\ \lambda]^T,\ c \geq 0, 0 \leq \lambda < 1.$$

The second method is to use the regularization method with multiple regularization matrix (11). The multiple regularization matrix is constructed based on the inverse of the TC kernel matrix which has closed-form expression:

$$D^{\mathrm{TC}}(\lambda) = (K^{\mathrm{TC}}(\eta))^{-1} \qquad (26)$$

where $K^{\mathrm{TC}}(\eta)$ is the kernel matrix constructed based on the TC kernel (25) with $c = 1$. From Carli et al. (2017) and Marconato et al. (2016), we know that D^{TC} has closed-form expression

$$D^{\mathrm{TC}}_{i,j}(\lambda) = \frac{\eta_{ij}}{1-\lambda}(-1)^{i+j}\lambda^{-\frac{i+j}{2}}\lambda^{\frac{|i-j|}{2}} \qquad (27)$$

where

$$\eta_{ij} = \begin{cases} 0 & \text{if } |i-j| > 1, \\ 1+\lambda & \text{if } i = j = 2, \ldots, n-1, \\ 1 & \textit{otherwise.} \end{cases} \qquad (28)$$

Moreover, we have $D^{\mathrm{TC}} = (F^{\mathrm{TC}})^T F^{\mathrm{TC}}$ with

$$F^{\mathrm{TC}}_{i,j}(\lambda) = \begin{cases} \lambda^{-i/2}(1-\lambda)^{-1/2}, & \text{for } j = i, i < n \\ -\lambda^{-i/2}(1-\lambda)^{-1/2}, & \text{for } j = i+1, i < n \\ \lambda^{-n/2} & \text{for } i = j = n \end{cases}.$$

Then we construct the multiple regularization matrix

$$D(\eta) = \sum_{i=1}^{15} \eta_i D_i = \sum_{i=1}^{15} \eta_i F_i^T F_i, \qquad (29)$$
$$\eta = [\,\eta_1\ \ldots\ \eta_{15}\,], \eta_i \geq 0, i = 1, \ldots, 15,$$

where for $i = 1, 2, \ldots, 15$,

$$D_i = D^{\mathrm{TC}}(\lambda_i), F_i = F^{\mathrm{TC}}(\lambda_i), \lambda_i = 0.85, 0.86, \cdots, 0.99.$$

For both methods, the empirical Bayes method is used to estimate the hyper-parameter η. The noise variance σ^2 is estimated as follows: an FIR model with order $n = 125$ is first estimated with the least squares method and the sample variance is then used as the estimate of σ^2. Similar to Chen et al. (2012), we choose not to use the unknown input $u(t)$ with $t = 0, -1, \ldots, 1-n$.

To measure the difference between the true impulse response of the test system and the regularized impulse response estimate, the following measure of fit is used:

$$\textit{fit} = 100\left(1 - \left[\frac{\sum_{k=1}^{125}|g_k^0 - \hat{g}_k|^2}{\sum_{k=1}^{125}|g_k^0 - \bar{g}^0|^2}\right]^{\frac{1}{2}}\right),\ \bar{g}^0 = \frac{1}{125}\sum_{k=1}^{n} g_k^0$$

where g_k^0 and $\hat{g}_k$ are the true impulse response and its estimate at the kth time instant, respectively.

5.3 Simulation Results

For the test systems and data-bank, the average fits for the two tested methods are summarized in the table

	TC	MReg
Avg. Fit	53.9	54.1

where "TC" denotes the kernel-based regularization method (9) with the TC kernel (25) and "MReg" denotes the regularization method (11) with multiple regularization matrix (29). The distribution of the model fits are shown in Fig. 1. As can be seen, the regularization method (11) with multiple regularization matrix (29) behaves similarly to the kernel-based regularization method (9) with the TC kernel (25), both in terms of the average accuracy and the robustness for the test systems and data-bank.

6. CONCLUSION

In this paper, we have considered the regularization method with the multiple regularization matrix and we

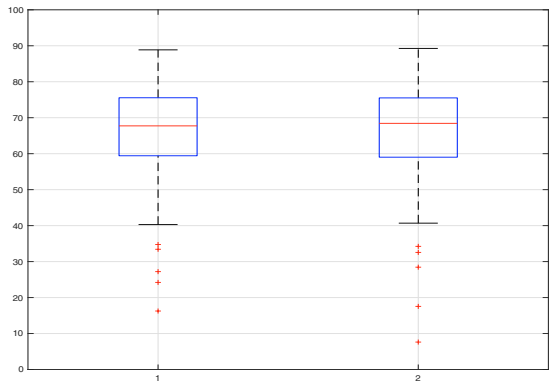

Fig. 1. Boxplot of the model fits for the two tested methods: the left column shows the result for the kernel-based regularization method (9) with the TC kernel (25) and the right column shows the result for the regularization method (11) with multiple regularization matrix (29). For both columns, there are 3 fits below zero, which are not shown for better display.

have shown that for the hyper-parameter estimation problem, the widely used empirical Bayes method can be expressed as a difference of convex programming problem, and hence a locally optimal solution could be found efficiently using sequential convex optimization techniques. In this preliminary work, we tested the multiple regularization matrix constructed based solely on the inverse of the tune-correlated kernel matrix. The simulation results showed the efficacy of the proposed method. In the near future, we will test the multiple regularization matrix constructed based on the filter-based regularization matrix introduced in Marconato et al. (2016). For example, to embed the prior knowledge that the system to be identified could be low-pass, high-pass, band-pass and etc., the fixed regularization matrix in the multiple regularization matrix can be chosen to be the filter-based regularization matrix with different orders, different cut-off frequencies, and different decay rate. Another topic to explore is the sparsity of the optimal hyper-parameters.

REFERENCES

Carli, F.P., Chen, T., and Ljung, L. (2017). Maximum entropy kernels for system identification. *IEEE Transactions on Automatic Control*, 62(3), 1471–1477.

Chen, T., Andersen, M.S., Ljung, L., Chiuso, A., and Pillonetto, G. (2014). System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques. *IEEE Transactions on Automatic Control*, (11), 2933–2945.

Chen, T., Ohlsson, H., and Ljung, L. (2012). On the estimation of transfer functions, regularizations and Gaussian processes - Revisited. *Automatica*, 48, 1525–1535.

Chen, T. (2018). On kernel design for regularized LTI system identification. *Automatica*, 90, 109–122.

Chen, T. (2019). Continuous-time DC kernel — a stable generalized first-order spline kernel. *IEEE Transactions on Automatic Control*.

Chen, T., Ardeshiri, T., Carli, F.P., Chiuso, A., Ljung, L., and Pillonetto, G. (2016). Maximum entropy properties of discrete-time first-order stable spline kernel. *Automatica*, 66, 34 – 38.

Chen, T. and Pillonetto, G. (2018). On the stability of reproducing kernel hilbert spaces of discrete-time impulse responses. *Automatica*.

Chiuso, A. (2016). Regularization and Bayesian learning in dynamical systems: Past, present and future. *Annual Reviews in Control*, 41, 24 – 38.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

Hong, S., Mu, B., Yin, F., Andersen, M.S., and Chen, T. (2018). Multiple kernel based regularized system identification with SURE hyper-parameter estimator. In *The 18th IFAC Symposium on System Identification (SYSID)*.

Hunter, D.R. and Lange, K. (2004). A tutorial on MM algorithms. *American Statistician*, 58, 30–37.

Latarie, J. and Chen, T. (2016). Transfer function and transient estimation by Gaussian process regression in frequency domain. *Automatica*, 52, 217–229.

Ljung, L. (1999). *System Identification - Theory for the User*. Prentice-Hall, Upper Saddle River, N.J., 2nd edition.

Marconato, A., Schoukens, M., and Schoukens, J. (2016). Filter-based regularisation for impulse response modelling. *IET Control Theory & Applications*, 11, 194–204.

Mu, B., Chen, T., and Ljung, L. (2018a). Asymptotic properties of generalized cross validation estimators for regularized system identification. In *The 18th IFAC Symposium on System Identification (SYSID)*.

Mu, B. and Chen, T. (2018). On input design for regularized LTI system identification: Power-constrained input. *Automatica, revised in January 2018, available from http://arxiv.org/abs/1708.05539*.

Mu, B., Chen, T., and Ljung, L. (2017). On the input design for kernel-based regularized LTI system identification: Power-constrained inputs. *Proc. 56th IEEE Conference on Decision and Control*.

Mu, B., Chen, T., and Ljung, L. (2018b). On asymptotic properties of hyperparameter estimators for kernel-based regularization methods. *Automatica*.

Pillonetto, G., Dinuzzo, F., Chen, T., De Nicolao, G., and Ljung, L. (2014). Kernel methods in system identification, machine learning and function estimation: A survey. *Automatica*, 50(3), 657–682.

Pillonetto, G. and Nicolao, G.D. (2010). A new kernel-based approach for linear system identification. *Automatica*, 46(1), 81–93.

Pillonetto, G. and Chiuso, A. (2015). Tuning complexity in regularized kernel-based regression and linear system identification: The robustness of the marginal likelihood estimator. *Automatica*, 58, 106–117.

Prando, G., Chiuso, A., and Pillonetto, G. (2017). Maximum entropy vector kernels for MIMO system identification. *Automatica*, 79, 326–339.

Yuille, A.L. and Rangarajan, A. (2002). The concave-convex procedure (CCCP). *Advances in Neural Information Processing Systems*, 2, 1033–1040.

Zorzi, M. and Chiuso, A. (2017). The harmonic analysis of kernel functions. *arXiv preprint arXiv:1703.05216*.